AD_____

Award Number: DAMD17-99-1-9174

TITLE: Computer Aid for the Decision to Biopsy Breast Lesions

PRINCIPAL INVESTIGATOR: Carey E. Floyd, Jr., Ph.D.

CONTRACTING ORGANIZATION: Duke University Medical Center
Durham, North Carolina 27710

REPORT DATE: July 2002

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are
those of the author(s) and should not be construed as an official
Department of the Army position, policy or decision unless so
designated by other documentation.

20030328 297

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE July 2002	3. REPORT TYPE AND DATES COVERED Final (1 Jul 99 – 30 Jun 02)	

4. TITLE AND SUBTITLE
Computer Aid for the Decision to Biopsy Breast Lesions

5. FUNDING NUMBERS
DAMD17-99-1-9174

6. AUTHOR(S):
Carey E. Floyd, Jr., Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University Medical Center
Durham, North Carolina 27710

E-MAIL:
cef@deckard.mc.duke.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Original contains color plates: All DTIC reproductions will be in black and white.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

The goal of this project is to improve the accuracy of the diagnosis of breast cancer from mammograms. In current practice, an estimated 2-10% of true cancers are not biopsied but are followed instead while between 60% and 90% of breast biopsies are performed on benign lesions. This report documents progress so far which demonstrates that this accuracy could be improved by a Case Based Reasoning approach to predict the outcome of a biopsy from the known biopsy outcomes for similar cases. The current version of the CBR performs with an accuracy of 61% on a retrospective set of consecutive cases for which the clinical diagnostic accuracy was 35%. This potential improvement in accuracy potentially improving the accuracy of the diagnosis of breast cancer from 35%to 61% for a set of 1023 cases. The CBR algorithm has been examined through evaluating and refining different techniques for the four fundamental tasks 1)specify a reference set of cases; 2)define a metric for the distance between cases; 3)define a rule (based on the distance metric) for selecting "similar" cases from the reference set; 4)specify a classification technique for predicting the outcome of biopsy from the known outcomes of the selected similar reference cases. In conclusion, CBR was found to be a promising technique for identifying benign cases.

14. SUBJECT TERMS
breast cancer computer-aided diagnosis

15. NUMBER OF PAGES
41

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

N/A   Where copyrighted material is quoted, permission has been obtained to use such material.

N/A   Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

N/A   Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A   In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

N/A   For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal law 45 CFR 46.

N/A   In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institute of Health.

N/A   In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA molecules.

N/A   In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

_____ 31 Oct 02_____
PI – Signature   Date

Table of Contents

4

Final Report of the Progress on Grant DAMD17-99-1-9174

For the Period of July 1999 to June 2002

Computer Aid for the Decision to Biopsy Breast Lesions

Hypothesis:

This work will test the hypothesis: "The results of breast biopsy can be accurately predicted from the results of biopsies for previous cases that had similar mammographic abnormalities."

Statement of Work

All of the tasks in the Statement of Work have been achieved. Tasks 9, 10, and 11 were replaced by similar tasks that were found to be more appropriate as a natural consequence of the discoveries made.

1   Obtain and setup computer equipment

    Performed in year 1

2   Program CBR software

    Performed in year 1

3   Examine ranking of features

Performed in year 1

4   Refine structure of database

Performed in year 1

5   Complete programming and test

Performed in year 1

6   Evaluate matching windows for continuous features

Performed in year 1

7   Determine appropriate evaluation (fitted or sampled ROC)

Examined in year 2

8   Evaluate matching rule (selection from 15 features)

Performed in year 2, refined through exhaustive search in year 3

9   Evaluate matching rule (weighted difference of features)

Replaced by Euclidean distance in year 3

10   Evaluate genetic algorithm for optimization

Replaced by exact search performed in year 3.

11   Evaluate matching rule based on likelihood of malignancy

Replaced by comparison of Hamming and Euclidean distance in year 3

Nomenclature

To clarify a potential source of confusion in this proposal, two terms are defined here: "feature" and "finding". The term "feature" refers to a variable while the term 'finding' refers to the value of a variable. For the categorical descriptions of mammographic abnormalities described below in Table 5, an example would be "the feature 'mass margin' has a finding of 'spiculated'.

Significance for reducing the number of benign biopsies

The lifetime risk of developing breast cancer has increased steadily from 1940, when the first statistics were collected, to the present risk of one woman in eight [1]. Several large studies have demonstrated that screening mammography can decrease the mortality due to breast cancer by 30%[2, 3]. Unfortunately, evaluating mammograms is a complicated task. Multiple radiographic features of each mammographic abnormality must be examined to determine whether further action such as follow-up or biopsy for histologic diagnosis is warranted. Although mammography is a sensitive tool for detecting breast cancer, the positive predictive value (PPV) has historically been low [4-6]. Due to the overlap of the radiographic appearance of benign and malignant breast lesions [6] as well as an overall conservative approach of physicians[7], only about 20% of women who undergo biopsy for mammographically suspicious non-palpable lesions have a malignancy by histologic diagnosis[5]. This relatively low Positive Predictive Value of mammography-induced biopsy is recognized as a significant problem. If the mammography screening recommendations of the American College of Radiology (ACR) and the American Cancer Society (ACS) are fully implemented, nearly all women over the age of 40 will undergo a yearly mammogram. Currently, the biopsy rate is 0.5 - 2.0% of all mammographic exams. Potentially, several million biopsies will be performed each year.[8] With the current accuracy, hundreds of thousands of women who do not have breast cancer would be

unnecessarily subjected to the discomfort, expense, potential complications, change in cosmetic appearance, and anxiety that can accompany breast biopsy.[5,9-11] In addition, the financial burden of these procedures (between $3500 and $5000 for excisional and between $1000 and $1500 for core biopsy) is substantial (around $100,000,000 per year)[5,8,9]. This project will develop an accurate computer-based system to provide a second opinion to assist the mammographer with the decision to biopsy.

The interpretation and decision process for a diagnostic mammogram is quite different from the screening mammogram. As a second reader in diagnostic mammography, the system could provide a mammographer with 1)a diagnosis, 2) an estimate of uncertainty for the diagnosis, and 3)sample images from the set that were accepted as similar. The mammographer can use this additional information for the decision to recommend biopsy or follow-up. A significant value to the clinician is that the decision aid potentially contains information derived from more cases than any mammographer could have ever seen and thus provides access to an experience base that would not otherwise be available.

The anticipated clinical impact of this CBR second opinion will be to increase the diagnostic accuracy of mammography for predicting malignancy of breast lesions. This will be achieved by decreasing the number of patients sent to biopsy with benign lesions and by decreasing the variability of diagnosis for mammography.

Year 1
Key Research Accomplishments

1   Implemented CBR algorithm on a Unix Workstation (SOW 1,2)

2   Feature selection was examined for Hamming distance (SOW 3,4, 6)

3   Initial performance evaluated (SOW 5)

8

To evaluate the contribution of individual findings, the performance of the algorithm was evaluated on a subset of all possible combinations of the six input features. The combinations that were tested are shown in table 1. These combinations represent the logical choices of grouping for these features. All eight feature combinations were examined and their performance was evaluated for a reasonable range of distance cut off values.

Table 2 Findings included in the matching rules

Findings	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
Age	X	X	X	X	X	X	X	X
Mass Margin	X	X	X	X	X	X	X	X
Calcification Description	X	X	X	X	X	X	X	X
Mass Density			X	X	X	X		
Calcification Distribution			X	X			X	X
Associated Findings		X		X		X		X

Table 1 The table shows which findings were included in each of the eight matching rules that were tested.

A receiver operating characteristic curve for the CBR performance is shown in fig. 1 below. Note the encouraging behavior at high sensitivity. The sensitivity remains very high as the false positive fraction (FPF) decreases and does not significantly decrease until the FPF has dropped to 0.6 (specificity of 0.4). With a threshold of 0.2, 126 benign biopsies could be avoided at a cost of 2 missed malignancies.

9

CBR: ROC result

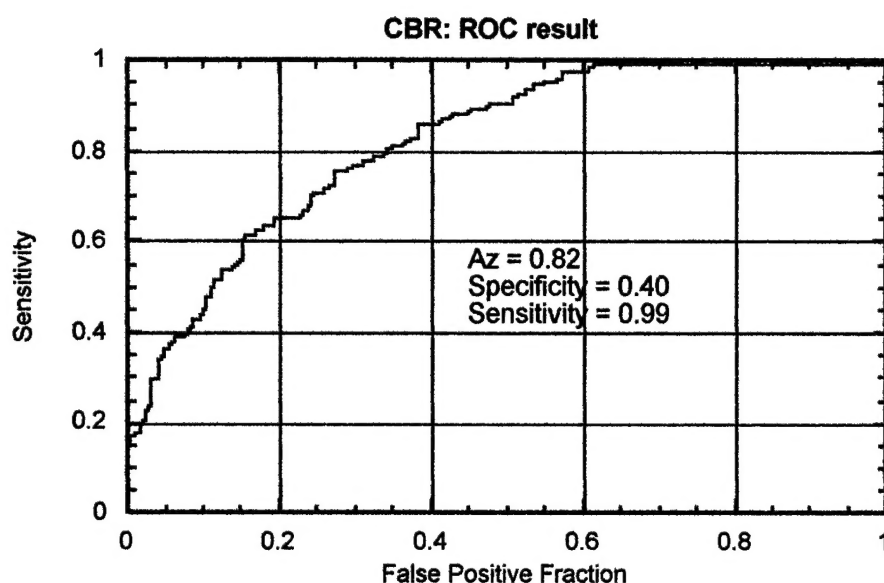Fig. 1. ROC plot of CBR output values for all benign and malignant cases.

Year 2

Key Research Accomplishments

1   Analyzed distribution of findings in the case database (SOW 4)

2   Non-parametric ROC evaluation of the classifier performance was performed (SOW 7)

3   Feature selection was examined for Hamming distance (SOW 8)

4   Hamming distance metric was evaluated (SOW 8)

A CBR system was developed to classify cases referred for biopsy. The CBR was evaluated on a set of 500 cases from Duke (described in more detail below) using round-robin sampling. All cases were referred to excisional biopsy and the truth for evaluating the classification of each case was abstracted from the pathology report. Of these 500 diagnostic mammography cases that were that were referred to biopsy, 326 (64%) were benign. While this fraction is higher than the value of 20% typically quoted as a national average, it is consistent with that seen at other teaching hospitals. In the framework of the specific aims of this proposal, the properties of this CBR include:

Table 1 Characteristics of CBR used in feasibility studies	
Reference data	500 Retrospective biopsy cases from Duke
Case Encoding	Uniformly scaled rank order
Similarity Metric	Hamming Distance
Similarity Selection	Threshold applied to Hamming distance metric
Classification metric	Probability of malignancy

Table 1: Characteristics that define the CBR.

Analyzed distribution of findings in the case database

Here we present some characteristics of the reference database that has been acquired. The database consisted of cases that were evaluated at diagnostic mammography after being called back due to an abnormality observed in a screening examination. All of the cases were non-palpable and were referred to biopsy. Cases were excluded if a previous biopsy or surgery had been performed at the site of the abnormality. Outcomes were established from the pathology report. Each case included 1) the

11

mammographers' description of the abnormality using the BI-RADS™ lexicon, 2) known epidemiological risk factors for breast cancer; and 3) outcomes in the form of benign or malignant status as determined by biopsy. The risk factors are routinely acquired by a short patient interview conducted by mammography technologists at the time of the diagnostic examination. Of the 500 lesions evaluated in the feasibility studies, there were 232 masses alone, 192 microcalcifications alone, and 29 combinations of masses and associated microcalcifications. The remaining 47 lesions included various combinations of architectural distortion, regions of asymmetric breast density, areas of focal asymmetric density, and areas of asymmetric breast tissue. Patient age ranged from 24 to 86 years with a mean value of 55 years. At biopsy, 326 (65%) of the lesions were found to be benign while 174 (35%) were found to be malignant. Currently (as of May 2001), our database contains around 1500 cases that were examined at diagnostic mammography and were referred to biopsy at Duke University Medical Center between 1992 and 2000. While this does not represent all of the consecutive cases, the omissions are believed to be random and these data are considered to represent an unbiased sample of the population of cases to which the decision system would be applied.
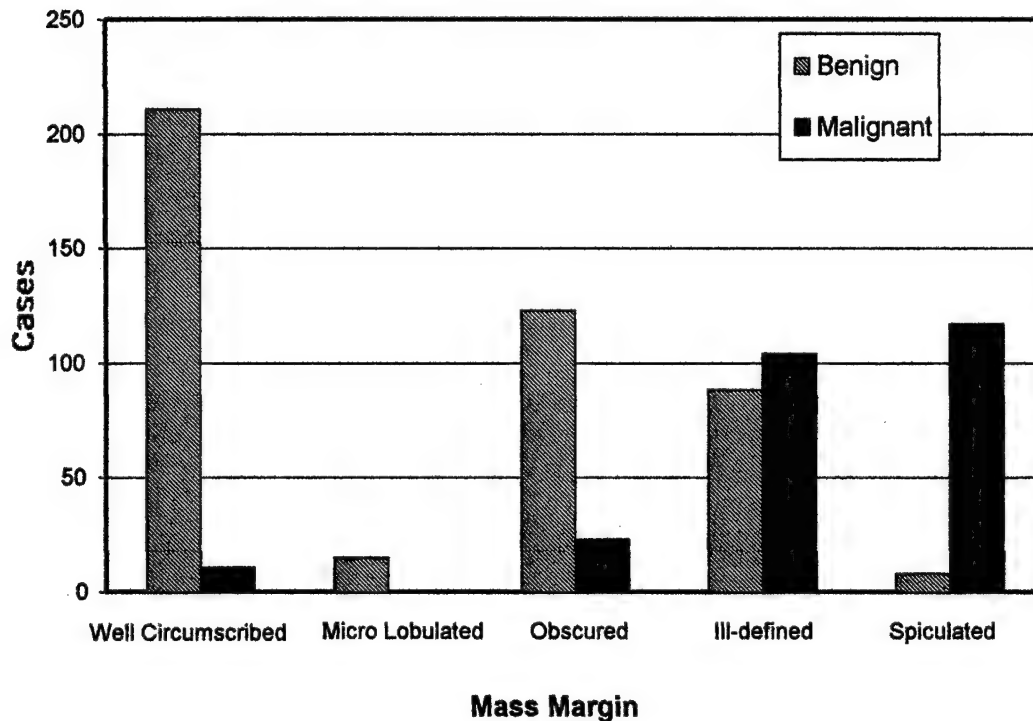
Distribution of Cases by Mass Margin

Fig. 2 The distribution of cases by mass margin is shown with malignant cases represented by the dark bars and the benign cases represented by the striped bars.

However, as combinations of features are considered, there is clear evidence that several of the features are not independent and that several of the joint probability distributions are not well determined. From our experience with the parametric fitting described above, we feel that it is important to avoid parametric assumptions where possible in this problem and propose to acquire more cases. When these joint distributions are examined from the data, numerous discontinuities are evident that are believed to be the result of too few cases. This is a concern for a CBR since these distributions are estimated directly from the reference data. As an example, consider the distribution of categorical findings for the mass margin and mass shape features shown in Fig. 2 and Fig. 3. Two observations are evident from Fig. 2. First, masses with Micro Lobulated margins are rarely referred to biopsy. Second, ignoring the Micro Lobulated category, the distributions for benign and malignant

13

cases are monotonicly decreasing/increasing respectively with the findings ordered as shown (which is consistent with the BI-RADS™ specification). From Fig. 2 there seem to be a sufficient number of cases to describe these distributions. In Fig.3 is the distribution of the Mass Shape feature. Here the distributions are not monotonic but the shape is still rather well defined although the relationship between the first three categories for benign masses is uncertain. When the dependence on mass shape is also considered, as shown in Table 2, it is clear that 1)these two features are not independent and 2)the form of the dependence is not well determined with the current number of cases, particularly for the benign masses.
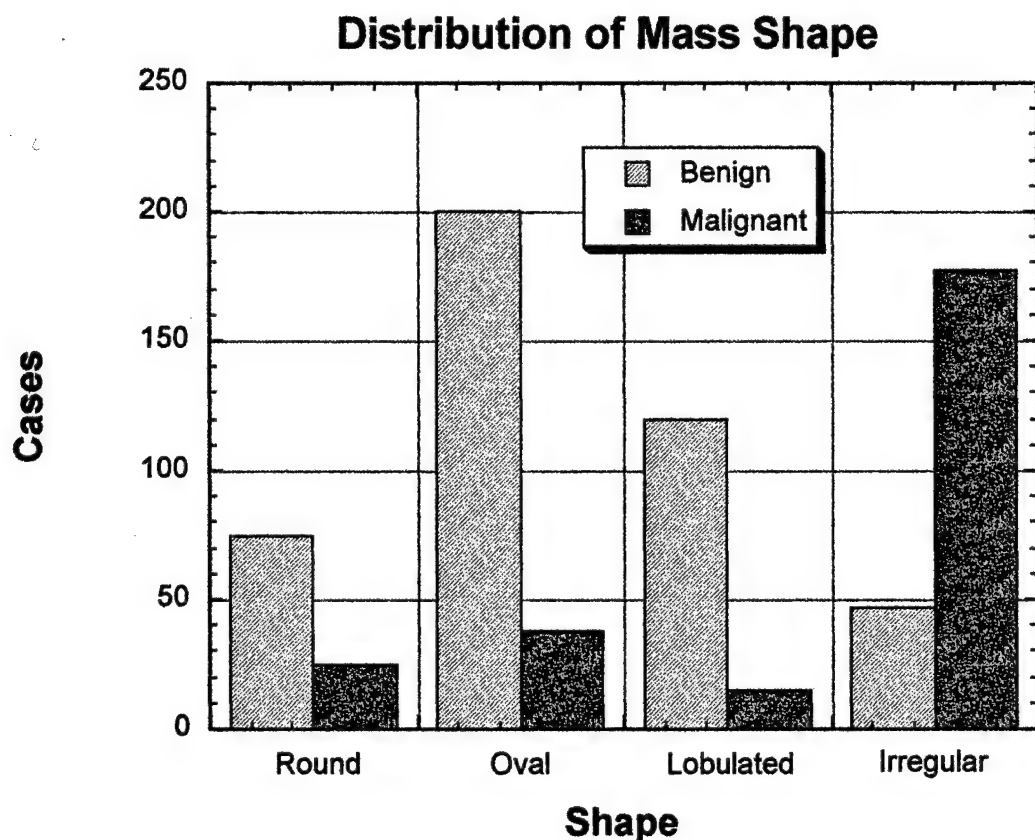
Distribution of Mass Shape

Fig. 3 The distribution of cases by mass shape is shown with malignant cases represented by the dark bars and the benign cases represented by the striped bars.

Shape	Margin	Well-	Circumscribed Micro-Lobulated	ill-defined	Obscured	Spiculated
Shape Benign	round	38	2	22	12	1
	oval	106	5	59	29	1
	lobulated	65	5	34	16	0
	irregular	1	2	7	31	6
Malignant	round	2	0	3	17	3
	oval	7	0	9	19	3
	lobulated	2	1	5	7	0
	irregular	0	0	6	61	110

Table 2 The joint distribution of cases by mass margin and mass shape is shown in this table with malignant cases in the lower pane and benign cases in the upper pane.

Define mathematical representation of a case

We have examined choices for the representation of the cases beginning with the representation used to develop our ANN classifiers. Cases are represented by a vector of features each of which has a number of possible categorical values or findings. BI-RADS™ was developed as a reporting lexicon and not as a direct indicator of probability for disease and while the assignment of numerical values to the categories is not provided, the lexicon does describe a rank order among many of the findings. From this in combination with discussions with several mammographers, a weighting (or value) scale has been developed and used successfully in the previous CBR and ANN analysis. These weights are

presented with the findings in Table 5 below. Values were assigned as normalized rank orderings of the categorical values in each finding independently and were intended to rank the possible descriptions in order of their likelihood of malignancy.

Feature selection was examined for Hamming distance

We examined the sensitivity of the CBR to the method used for selecting cases. The selection rule is a combination of a distance metric and a threshold technique. Here several sets of features were examined for computing the Hamming distance and the cutoff threshold was varied. Of interest is the observation that performance increased when the distance increased from 0 (which required an exact match) to 1 (which allowed one of the 6 features to differ). The best performance was found when only three features were required and up to one was allowed to differ. We believe that better performance will be obtained with more than three features but that this will require more cases. This seems likely when considering that with these three features: Mass Margin, Calcification Description and Age, only cases with calcified masses (10% of the cases) could possibly non-null findings for all three features. As a side note, while the best CBR performance is slightly less than the best ANN performance on these cases, the ANN performance is close to chance if only three features are provided.

Hamming distance metric was evaluated

Table 3 Case Based Reasoning: Performance for Hamming Distance						
Number of Features	Feature set	Distance Threshold	ROC Area	Partial ROC Area	Specificity at 100% Sensitivity	Specificity at 98% Sensitivity

6	A	0	0.70	<0.05	<0.01	<0.01
6	A	1	0.79	0.2	<0.01	<0.01
3	B	1	0.83	0.45	0.25	0.41

Table 3. Performance of CBR with Hamming distance as a function of distance threshold and features sets Feature set A: Age, Mass Margin, Mass Shape, Calcification Description, Calcification Distribution, Associated Findings; set B: Age, Mass Margin, Calcification Description.

Table 4 Performance for different thresholds on the probability of malignancy					
Probability Threshold	Sensitivity (%)	Specificity (%)	Positive Predictive Value (%)	Benign Biopsies Avoided	Malignancies Missed
Accept All Cases	100	0	35	0	0
0.10	100	25	42	81	0
0.21	98	41	46	134	10

Table 4. Performance of Case Based Reasoning System for different thresholds applied to the predicted probability of malignancy.

Table 5 - Input Features for breast biopsy cases

BI-RADS™ Lesion Descriptors				**BI-RADS™ Lesion Descriptors**			
Input Node	Feature	Finding	Value	Input Node	Feature	Finding	Value
1	Calcification Distribution			8	Location		
		no calcifications	0			___o'clock	
		diffuse	0.2			axillary tail	0

	regional	0.4		posterior	0.2
	segmental	0.6		middle	0.4
	linear	0.8		anterior	0.6
	clustered	1.0		subareolar	0.8
				central	1.0

2	Calcification Number			9	Associated Findings	
	no calcifications	0				
	< 5	0.33			none	0.00
	5 to 10	0.66			skin lesion	0.13
	> 10	1.0			hematoma	0.25
					trabecular thickening	0.38
3	Calcification Description				nipple retraction	0.50
	no calcifications	0			skin retraction	0.63
Benign-like findings	milk of calcium-like	0.2			skin thickening	0.75
	rim	0.2			architectural distortion	0.88
	skin	0.2			axillary adenopathy	1.00
	vascular	0.2				
	spherical	0.2	10	Special Cases		
	suture	0.2			none	0
	coarse	0.2			intramammary lymph node	0.25
	large rod-like	0.2			asymmetric breast tissue	0.5
	round	0.2			focal asymmetric density	0.75
other	dystrophic	0.4			tubular density or solitary dilated duct	1.0
	punctate	0.6				
	indistinct	0.8				
	pleomorphic	0.9				
	fine branching	1.0				

Features Involving Personal and Family History

Input Node	Feature	Finding	Value
11	Age		in years
12	Personal History	none	0
	of breast cancer	positive	1

4	Mass Margin	
	no mass	0
	well circumscribed	0.2
	microlobulated	0.4
	obscured	0.6

#	Feature	Finding	Value	#	Feature	Finding	Value
		ill-defined	0.8	13	History of Prior Ipsilateral Benign Biopsy	none	0
		spiculated	1.0			positive	1
5	Mass Shape			14	Family History of breast cancer	none	0
						positive	1
		no mass	0				
		round	0.25	15	Menstrual History	pre-menopausal	0
		oval	0.5			post-menopausal	1
		lobulated	0.75				
		irregular	1.0	16	Estrogen/Progesterone Therapy	none	0
						positive	1
6	Mass Density						
		no mass	0				
		fat-containing	0.25				
		low density	0.5				
		isodense	0.75				
		high density	1.0				
7	Mass Size		mm				

Table 5 shows the case representation that was evaluated. The "value" shown indicates the quantitative values assigned to individual findings in the preliminary data. These were initially assigned by uniformly distributing the rank-ordered findings between 0 and 1 for each feature.

Non-parametric ROC evaluation of the classifier performance was performed

Typically, published ROC curves are smooth since they are obtained though a parametric representation of the data. For a five-category human observer response experiment, this parameterization is necessary and is usually obtained using the software developed by Dr. Charles Metz of the University of Chicago. In the initial experiments, we found that the fitted curves did not accurately follow our data in the regions of high sensitivity which is exactly where we have the most interest in comparing techniques. Outputs of this CBR, the histogram of the negative cases followed a

distribution that did not appear to be normal. .After consulting with Dr. Metz, we decided that a non-parametric evaluation of the ROC performance would be more appropriate for these data. The source of our difficulty lay in the deviations from the normal distribution that are found in the tails of the probability density functions from the CBR. Interestingly, the ROC area estimates agreed very well, but the shapes were different. For this reason, all ROC curves are presented in a non-parametric form. That is, they are plotted from the data rather than from a fit to the data. With 500 or more continuous valued outputs, the Trapezoid Rule for computing the area gives sufficient accuracy. A convenience of the parametric fitting software is that they provide an estimate of the significance of any differences in performance for paired data. To estimate the significance of a difference computed with non-parametric methods, the mean values and variances (including covariances) for all performance measures were obtained by bootstrap sampling[49]. For the results presented here, 3000 samples were found to provide asymptotically stable estimates for all performance measurements.

Performance was evaluated by the receiver operating characteristic curve (ROC), the Partial Area Index ($_{0.90}A_z$) computed as the ROC area (scaled by 10) for sensitivities greater than 90%, and the specificity at 98% sensitivity. The ROC curve is shown in Fig. 4 below. Particularly encouraging is the behavior of the curve at high sensitivity, seen more clearly in the plot of $_{0.90}A_z$ in Fig. 5. The sensitivity remains very high as the false positive fraction (FPF) decreases. The sensitivity does not decrease below 98% until the FPF has dropped to 0.6 (specificity of 0.4). At this operating point, 130 of the 326 benign biopsies could be avoided with delayed diagnosis for only two malignant cases.
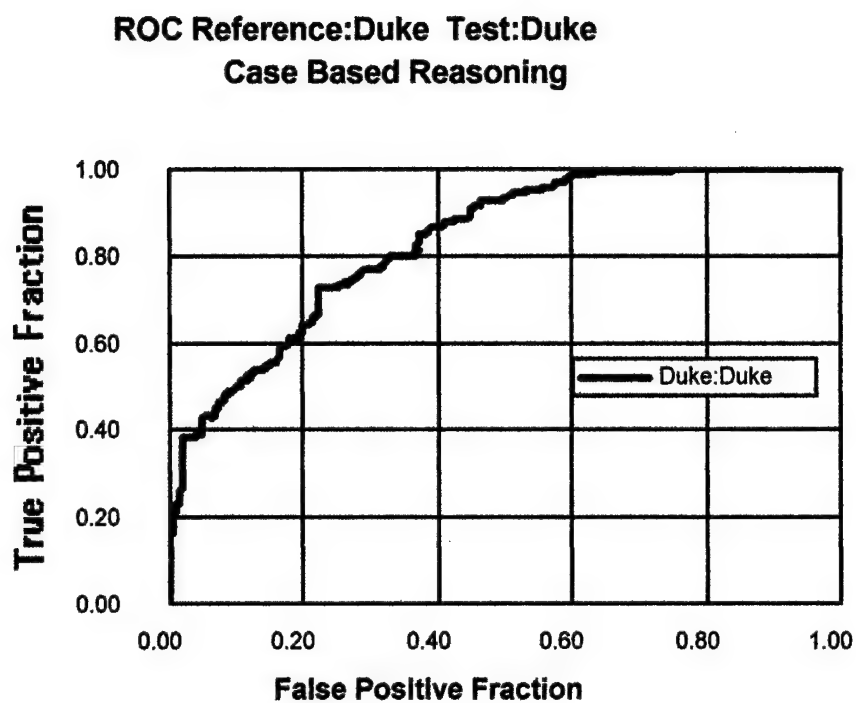
ROC Reference:Duke Test:Duke
Case Based Reasoning

Fig. 4. Full ROC curve for the CBR described in Table 1

Partial ROC Reference:Duke Test:Duke
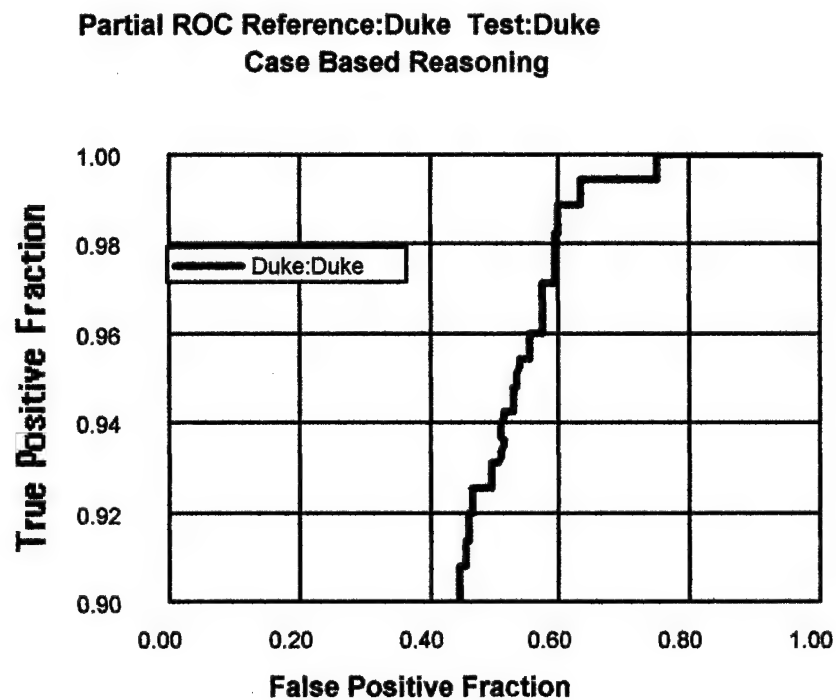**Case Based Reasoning**

Fig. 5. Partial ROC curve for the CBR described in Table 1. This performance measure is of more

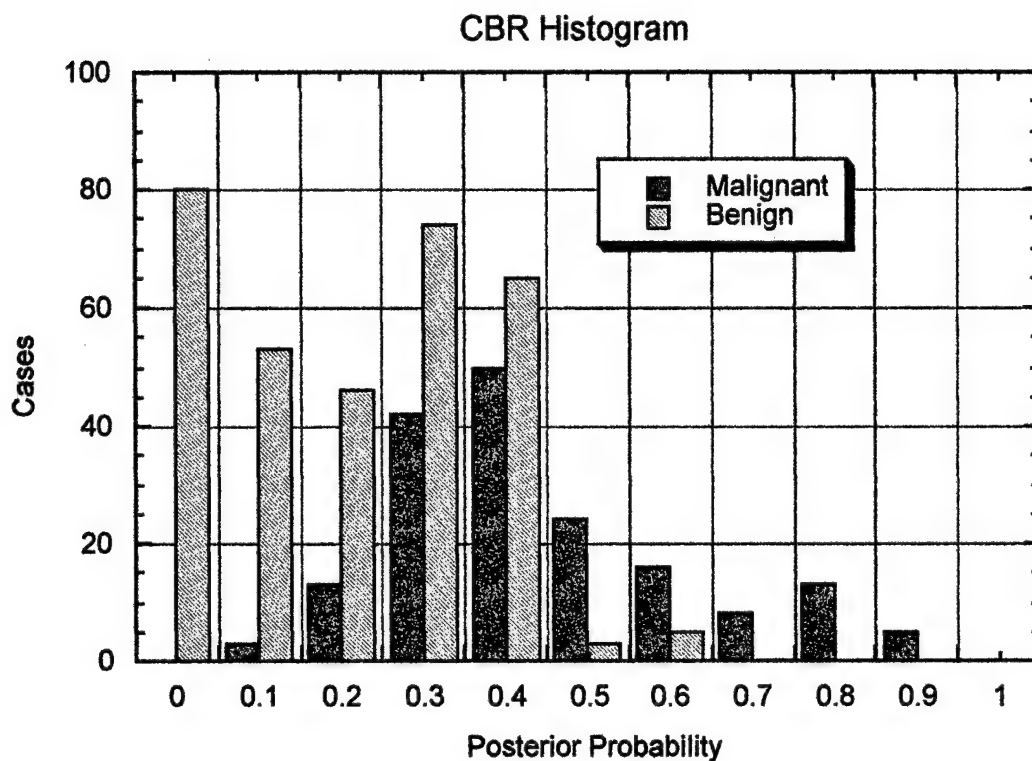clinical relevance than the full ROC for this cancer diagnosis task.

CBR Histogram

Fig. 6 The histogram of benign and malignant cases for the full range of the CBR output. The benign cases are represented by the striped open bars while the malignant test cases are represented by the gray bars.
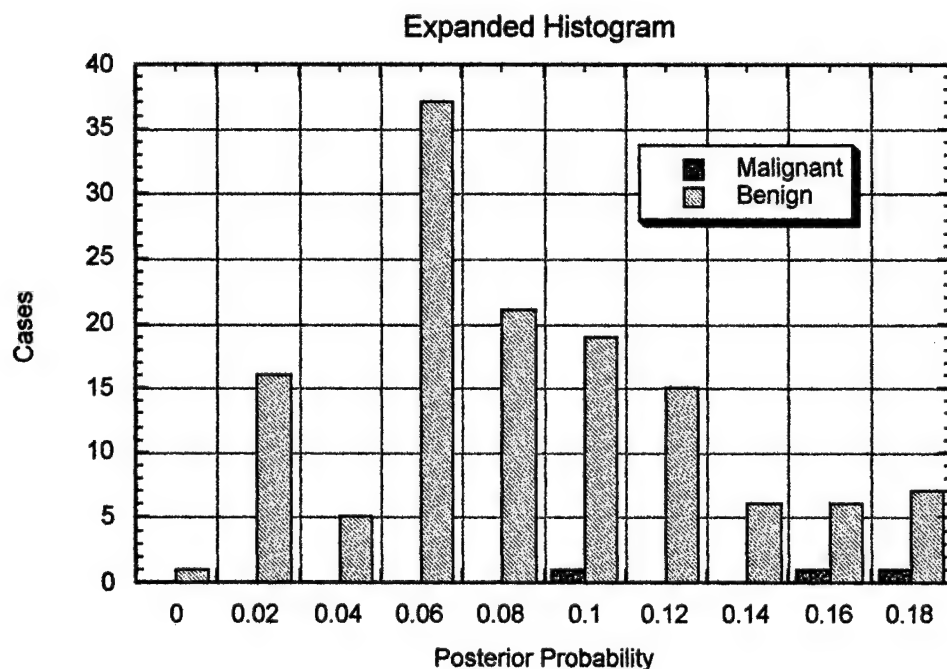
Expanded Histogram

Fig. 7 The histogram of benign and malignant cases for an expanded region of low probability for malignancy.

As seen in Fig. 6, moderate separation of the benign and malignant cases was achieved resulting in an ROC area of 0.83. Since there are some benign cases to the left of all malignant cases, biopsy could be avoided for these without missing any of the malignancies. To further examine this region, the histogram is expanded in Fig. 7 for the region assigned low probability of malignancy.

In this low probability region, there are 133 benign cases and only three malignant cases. The benign cases are represented by the striped open bars while the malignant test cases are represented by the gray bars.

The portion of the ROC curve that is of greatest interest is the region of greatest true-positive fraction (i.e. highest sensitivity) since few radiologists or patients would be willing to miss a larger fraction of breast cancers for the sake of high specificity. The cases populating this region are those that were assigned the lowest probability of malignancy.

It is interesting to note that the cancer shown farthest to the left in Fig. 7 is a 45 year old woman with a small well-circumscribed mass. These characteristics all would indicate a benign mass and the CBR agreed. The critical information that was not included in the model is that this mass was not seen in a previous mammogram. This information will be included in the proposed studies. In addition, it is interesting to note the features of the benign lesions that were assigned a probability lower than any of the malignancies. These are all masses and include 60 with well circumscribed margins, and one mass with a well circumscribed margin and with associated calcifications described as indistinct, one mass with a microlobulated margin, 18 masses with obscured margins, and one mass with an ill-defined margin and with associated calcifications described as coarse.

At sensitivity of 0.98 (relative to all biopsied lesions) the specificity is 0.4. Thus, 40% of the benign biopsies could have been avoided at the cost of delaying diagnosis for 2% of the malignancies. The positive predictive value for these data would be increased from 35% to 46%. These results demonstrate feasibility for developing CBR as a decision aid for breast biopsy using the BI-RADS™ lexicon to index the cases.

Year 3
Key Research Accomplishments

5   Expanded database
6   Performed exhaustive feature selection (SOW 10)
7   Evaluated Hamming and Euclidean distance measures (SOW 9)
8   Compared performance (SOW 11)

24

Expanded Database

Our current database consists of biopsy cases from three medical centers. We have approximately 1530 previously collected biopsy cases from Duke University Medical Center (Duke), and 1000 cases from the University of Pennsylvania Health System (Penn).[58] We have also extracted approximately 1979 cases from the publicly-available Digital Database for Screening Mammography[59] (DDSM) from the University of South Florida (USF). The DDSM database contains cases from multiple institutions, although the majority of the cases come from Massachusetts General Hospital and Wake Forest University School of Medicine. The cases were collected between 1988 and 1999. The Duke database, collected between 1991 and 2000, consists of consecutive core- and excisional-biopsy cases from Duke University. The Penn database is a collection of consecutive excisional biopsy cases from Penn, collected between 1990 and 1997. The cases were collected as part of standard clinical practice. The data for each case was compiled either at the time of the decision to biopsy, or retrospectively while blinded to biopsy outcome. Each suspicious mammographic lesion was described by a dedicated breast-imaging radiologist. Biopsy outcome for each case was obtained from the histopathological analysis.

Evaluated Hamming and Euclidean distance measures

Similarity between cases can be computed over the vector of findings for each case, resulting in a distance between two cases. Thresholding this distance establishes whether any two cases are similar or not. When this is performed between a test case and the reference database, the process can be visualized as a nearest-neighbor method of PDF estimation. Nearest-neighbor estimation smoothes the feature distributions, since each reference case can be reused several times as a nearest neighbor to any given test case. There are several distance measures that can be used; some are computationally simple

and some are more robust statistically. The distance measures we will examine include Mahalanobis distance,[63] Euclidean distance,[63] Hausdorff distance,[64] and Hamming distance.[63] The Hamming distance is calculated as the number of features that differ in value between two cases. For every feature, if the value differs between the two cases, the Hamming distance is increased by one. If the difference between the patient age of two cases differs by more than 3 years, the distance is also increased by one. The age difference of 3 was found to be locally optimal over the range 1 to 10 with all other parameters fixed. For n features (and thus a vector of length n representing a case), the Hamming distance between two cases can range from 0 (no features differ) to n (all features differ). For example, for ten features the Hamming distance between two cases can range from 0 to 10. Formally, the Hamming distance between the test case and the reference case, is

$$D_{Hamming}(test, ref) = \sum_{i=1}^{n} (1 - \delta(\phi_{i_{test}} - \phi_{i_{ref}}))$$

where ϕ_i test is the feature value for the specific case (test or reference) and feature, and m is the number of features used. The δ function returns 1 if the test case and the reference case feature values are the same and 0 if they are different. As mentioned before, if the ages differ by less than 3 years, they are considered the same. As stated in the beginning, the Hamming distance is the number of features that differ in value between the two cases. The Euclidean distance can be also implemented for measuring the distance between feature values (findings) of two cases. The Euclidean distance is computed as the square root of the sum of the squared differences between corresponding findings of two cases. For Euclidean distance measure, each finding is also normalized using linear scaling to unit

range. For a lowerbound value of ϕ_L and an upperbound value of ϕ_U for the specific feature, the new feature value would be

$$\phi = \frac{\phi_{original} - \phi_L}{\phi_U - \phi_L}$$

Therefore, after normalization, each feature value is in the range [0,1]. The Euclidean distance between the test case and a reference case is,

$$D_{Euclidean}(test, ref) = \sqrt{\sum_{i=1}^{n} \left| \phi_{i_{test}} - \phi_{i_{ref}} \right|^2}$$

where n is the number of features (2 in this setup), and ϕ is the normalized feature value for the specific case (test or reference), and feature i. The Euclidean distance is unaffected by translations and rotations of the feature space, but is sensitive to linear transformations. One way to avoid this problem is to apply standardization. Standardization transforms the features to have zero mean and unit variance. Standardization can prevent certain features from dominating distance calculation because they have large numerical values. However, if the spread of the features is due to difference in classes (benign vs. malignant), standardization can be undesirable.

Feature selection

The performance of a classifier can be significantly diminished by using too many input features; the need for the reduction of the number of features for any CAD system is well recognized.[71] The number of training data points (cases) for a classifier needs to be an exponential function of the feature

dimension.[72] This behavior is referred to as a "curse of dimensionality," and can lead to the "peaking phenomena." In the "peaking phenomena," additional features can degrade the performance of a classifier, if the number of training cases is small compared to the number of features.[72] Therefore, one of the main goals of this project is to find and present the most influential feature subset for improved, robust and consistent performance of the CBLR system. Furthermore, feature PDF estimation (one of the major steps in our classifier) becomes increasingly complicated with increasing dimensionality of the feature space. Therefore, we will investigate how feature selection methods can reduce the dimensionality of our feature space for the CBLR. Feature selection can be defined as selecting a subset of size m from a set of d features, that leads to the largest value of some criterion function $J(\cdot)$.[72] In order to select features which are the strongest and most consistent performers, we can perform the optimal exhaustive sequential search[72] of the feature space. This represents the most direct approach to feature selection - examining all (d_m) subsets of size m. Since we also examine all possible m features, this means that $2_d - 1$ feature combinations need to be examined. The number of combinations grows exponentially with d. For d=10, excluding the empty set (no features) means that 1022 ($2_{10} - 2$) feature combinations are examined in all. This brute force approach allows for a thorough examination in the search for best performance, yet in our case is still feasible computationally. Since feature selection is performed during the development of the classifier (rather than during use – testing), the computational speed of the feature-choosing algorithm is less important than optimality. The criterion function $J(\cdot)$ is chosen to be $_{0.90}AUC$, since we are most interested in reducing the number of benign biopsies while missing a minimal number of malignancies (and thus in the high sensitivity region of the ROC curve). Although the straightforward approach to exhaustive feature selection requires selecting the feature subset with the largest value of some criterion function $J(\cdot)$,[72] we have found that there were several feature combinations with very high and similar values of $J(\cdot) = _{0.90}AUC$, yet containing different features. The best single feature combination (strategy) could

thus be a minor stroke of luck, include a feature that does not really supply relevant information, and actually decrease performance if and when new cases are added to the database. Therefore, we have sought an approach to choose features that occur most often in the strategies with highest values of $_{0.90}AUC$ as the best feature subset. To our knowledge, this is a novel approach to exhaustive feature selection. By choosing features that occur most often, we are increasing the chance that the performance will remain the same when our data set increases or changes in the future. Previous feature selection techniques in CAD mammography included methods such as genetic algorithms for mass features,[73] and stepwise linear discriminant analysis for calcification classification.[74-76] However, most feature selection methods do not usually exhaust all the available feature combinations in the search for the best subset, and are not optimal.[72] In contrast, we can apply the optimal exhaustive search of our feature space due to our relatively small (10) feature subset, and fast computational speed of contemporary computers.

Description of input data. The database used in this study consisted of 1433 biopsy-proven cases from Duke University Medical Center, described using ten features as listed in Table 1. Histopathological analysis of these cases found 502 malignancies, while 931 of the suspicious lesions were found to be benign. This suggests that 931 patients (65%) could have perhaps avoided the biopsy procedure. Please note that all of the cases in our database were deemed suspicious enough by dedicated breast imagers such that a biopsy was performed on each one.

Two distance measures for nearest-neighbor PDF estimation were examined, Euclidean and Hamming. (This corresponds to classic CBR, since distance measures are used for similarity computations, resulting in nearest-neighbor PDF estimation.) Exhaustive feature selection was performed. For ten features, this means that 1022 feature combinations (strategies) were examined. Top strategies were

29

defined as strategies with highest $_{0.90}$AUC, and the top 10 strategies were the 10 strategies rank-ordered by $_{0.90}$AUC. The features that occurred most frequently (60% or more) in the top 10 strategies for each distance measure were selected as the most consistent performers for that distance measure. In order to examine the dependence of the selected features on the distance measure, the best performing feature subset found using the Hamming distance measure was evaluated with the Euclidean distance measure, and vice versa. We shall refer to the best feature subset found and tested using the Hamming Distance measure as Highest Consistent Hamming strategy. We shall refer to the tested Euclidean strategy with feature subset from Hamming as Euclidean Corresponding Strategy.

Results - CBR with Hamming Distance.

Figure 8 shows the maximum value of $_{0.90}$AUC achieved for each possible m (m=1..10) features using Hamming distance measure. This was done by first selecting the best value of $_{0.90}$AUC for each of the 1022 feature combinations, and then grouping feature combinations with the same number of features m and choosing the best value for each group (10 groups). This means that for m=1 to m=9 several feature combinations were used for findings the maximum value, while for m=10 only one strategy was included. Please note that for each feature combination all similarity thresholds were used to find the maximum value of $_{0.90}$AUC. The maximum value of $_{0.90}$AUC occurred at m=6 features for the Hamming distance measure. The Hamming distance measure had lower values of $_{0.90}$AUC than the Euclidean distance measure over all possible m.
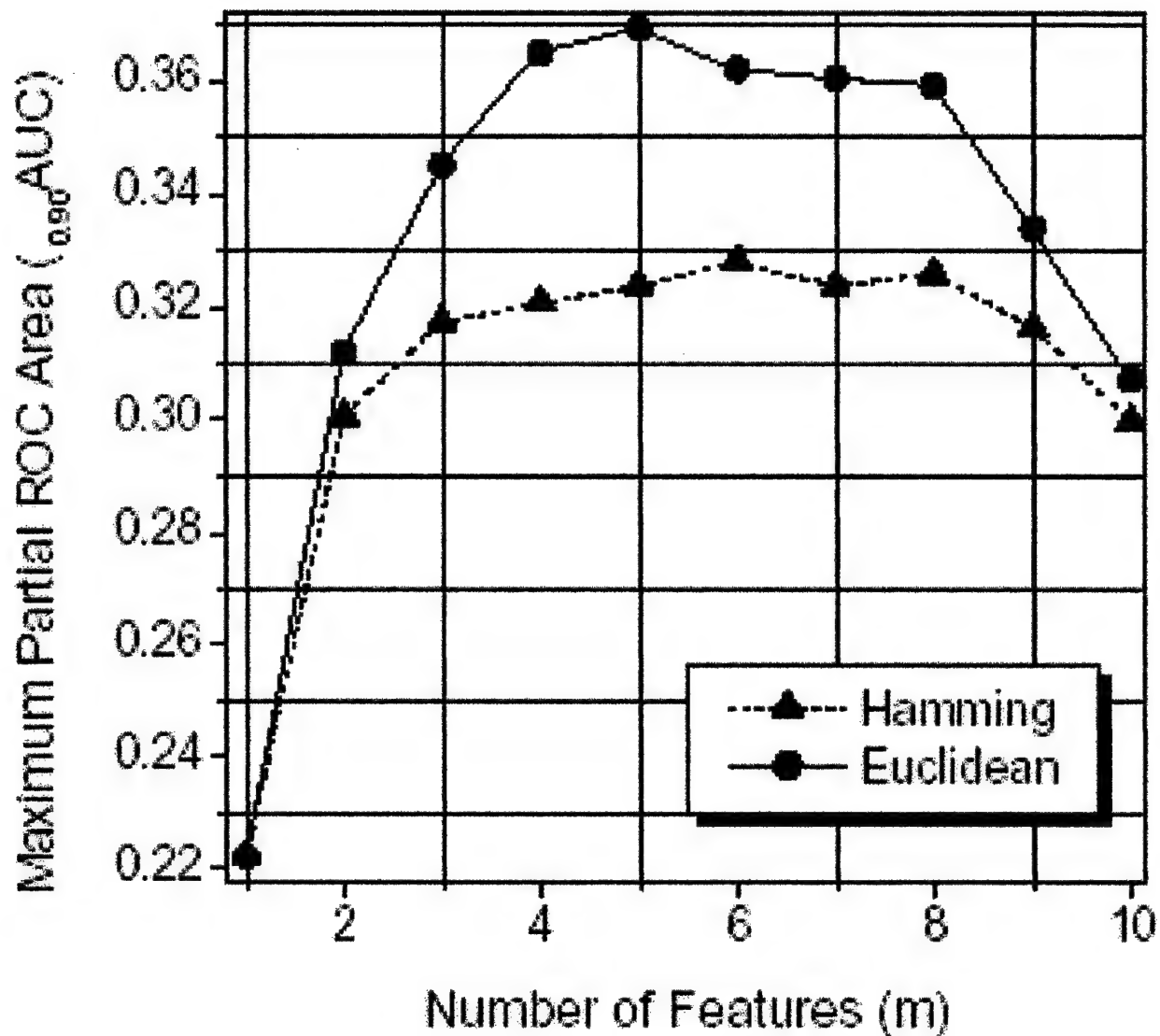
30

Figure 8: The maximum value of 0.90AUC achieved for each possible m features. Please note that the

0.90AUC using all features (for both distance measures) is lower than using only some of the features.

The single highest AUC from a single strategy using the Hamming distance measure had a value of

0.80. The highest individual $_{0.90}$AUC from the top $_{0.90}$AUC strategy was 0.33. Choosing the individual

best $_{0.90}$AUC strategy would spare 288 benign cases at 95% sensitivity, and 251 benign cases at 98%

sensitivity. (At 98% sensitivity, we are misclassifying 2% or 11/501 of the malignancies). The features used by the highest individual strategy are shown in Table 6 (rank 1).

Table 6: Example of the five highest strategies by 0.90AUC using the Hamming distance measure. Note thatthe values of 0.90AUC are practically the same, yet the features used are different in each case.

Rank	Features Used	0.90AUC	AUC
1	Age, Associated Findings, Mass Margin, Calcification Morphology, Mass Shape, Calcification Number	0.328	0.779
2	Age, Associated Findings, Mass Margin, Calcification Morphology, Mass Size, Calcification Distribution	0.327	0.791
3	Age, Associated Findings, Mass Margin, Calcification Morphology, Mass Size, Mass Shape, Calcification Number, Calcification Distribution	0.326	0.779
4	Age, Associated Findings, Mass Margin, Calcification Morphology, Mass Shape, Calcification Number, Calcification Distribution	0.324	0.787
5	Age, Mass Margin, Calcification Morphology, Mass Shape, Calcification Number	0.324	0.784

The best features for CBR with Hamming distance were found to be as follows. Three features - age, mass margin, and calcification morphology - occurred in all of the top 10 strategies for CBR using Hamming distance. Mass shape, calcification distribution, and calcification number occurred in a majority (7 or more) of the top 10 strategies. The other features occurred in less than 5 of the top 10 strategies. Mass density did not occur at all. Therefore, the features that were the most consistent performers for Hamming were chosen to be age, mass margin, calcification morphology, mass shape, calcification distribution and calcification number The CBR performance was then evaluated using these features and the Hamming distance measure, and the results (bootstrap averages) are presented in Table 7 as Highest Consistent Hamming Strategy. Since the feature selection, in effect, was done over all cases and the performance was evaluated on the original data set, this test set it not truly independent. The best performance of Hamming Highest Consistent strategy occurred at DTH_SIM_HAM of 2, meaning that cases were judged similar if 2 or fewer of the 6 features differed. Figure 9 shows the full and partial ROC curve for the Highest Consistent Hamming strategy from the Round Robin evaluation. Using the Hamming Highest Consistent strategy at 98% sensitivity
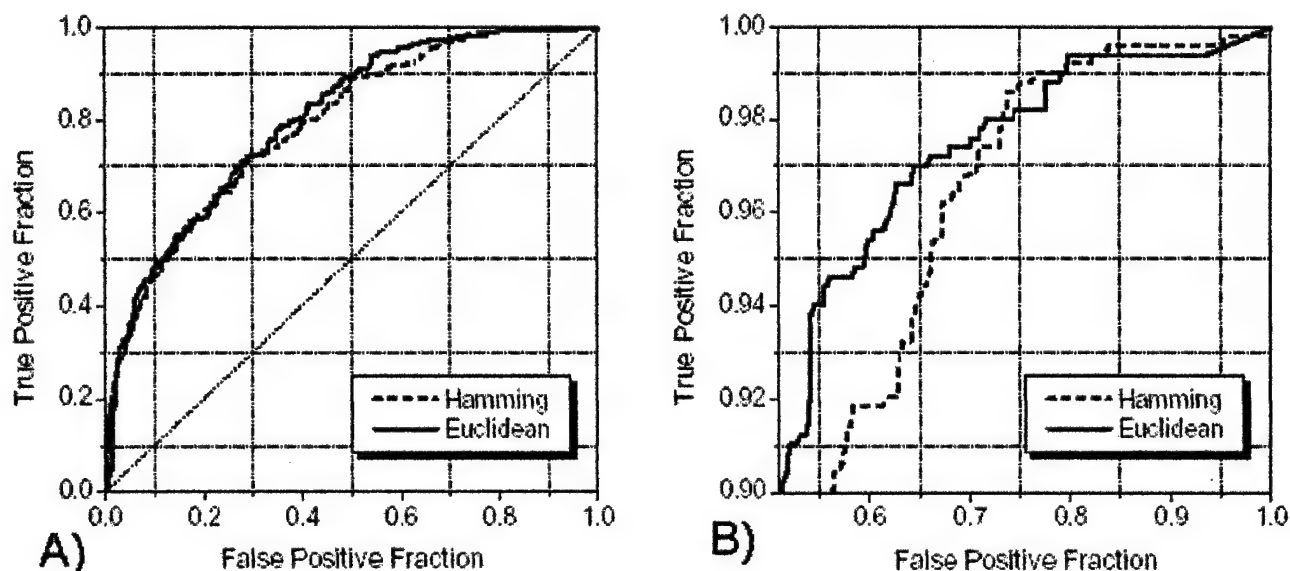
33

Compared performance

Figure 9: A) ROC curves for Highest Consistent strategies of the CBR classifier using Hamming and Euclidean distance measures. The thin dotted line extending from (0,0) to (1,1) represents chance behavior. B) Partial ROC curves for Highest Consistent strategies of the CBR classifier using Hamming and Euclidean distance measures. There is an almost statistically significant difference between the two curves (p=0.06). Note that while the area (0.90AUC) under the Euclidean curve is greater, the Hamming distance measure performs comparably at a sensitivity (true positive fraction) of 98%. (while misclassifying 2% or 11 of the malignancies) could potentially spare 251 (27%) of the benign cases. This represents an improvement from the clinical positive predictive value (PPV) of 35% to 42%. At 95% sensitivity (while misclassifying 5% or 26 malignant cases) the CBR could potentially spare from biopsy 316 (34%) benign cases. This raises the PPV to 44%.

Results - CBR with Euclidean Distance.

The highest AUC from a single strategy using the Euclidean distance measure had a value of 0.81. The highest individual 0.90AUC was 0.37. Choosing the individual best 0.90AUC strategy would spare 376

34

benign cases at 95% sensitivity, and 263 benign cases at 98% sensitivity. From Figure 8, the best single $_{0.90}$AUC using the Euclidean distance measure occurred at m=5 features. The best features for Euclidean distance CBR were found as follows. Age, mass margin, and calcification morphology occurred in all top 10 strategies - same as for Hamming. Unlike in Hamming, associated findings also occurred in all of the top 10 strategies, while mass density occurred 60% of the time. Special findings, mass shape, calcification distribution, and calcification number did not occur at all. Mass size occurred in 40% of the top 10 strategies. Therefore, the features that were most consistent performers for Euclidean were chosen to be age, mass margin, calcification morphology, associated findings, and mass density. The CBR system was next evaluated using these features and Euclidean distance measure, and the exhaustive search result of $DTH\_SIM\_EUC = 0.23$. The results (bootstrap averages) are presented in Table 7 as Highest Consistent Euclidean Strategy.

Table 7. Results from the CBR classifier for four strategies.

Distance Measure	Type of Strategy	AUC ± STD	$_{0.90}$ AUC ± STD	False Positive Fraction at 95% Sensitivity ± STD	Number of Cases Spared at 95% Sensitivity	False Positive Fraction at 98% Sensitivity ± STD	Number of Cases Spared at 98% Sensitivity
Hamming	Highest Consistent	0.79 ± 0.01	0.33 ± 0.03	0.66 ± 0.02	316 (34%)	0.73 ± 0.02	251 (27%)
Euclidean	Highest Consistent	0.80 ± 0.01	0.37 ± 0.03	0.59 ± 0.03	386 (41%)	0.73 ± 0.05	255 (27%)
Euclidean	Corresponding	0.79 ± 0.01	0.32 ± 0.03	0.64 ± 0.03	332 (36%)	0.75 ± 0.05	228(25%)
Hamming	Corresponding	0.80 ± 0.01	0.30 ± 0.03	0.66 ± 0.02	313 (34%)	0.77 ± 0.05	214(23%)

Table 7 Results from the CBR for four strategies.

Figure 9 shows the full and partial ROC curves for the Highest Consistent Euclidean strategy from the Round Robin evaluation. Using the Euclidean Highest Consistent strategy at 98% sensitivity could potentially spare 255 (27%) of the benign cases. This represents an improvement from the original PPV of 35% to 42%. At 95% sensitivity, approximately 386 (41%) benign cases could potentially be spared. This raises the PPV to 47%. For comparison, since every case in this database was referred to

35

biopsy, we can say that the original sensitivity was 100% (all malignancies were detected), while specificity was 0% (no benign lesions were spared) over the cases in our database. (Note that this is not representative of the radiologists' performance over a general screening or diagnostic population.). The same features were used with the Hamming distance measure, and the results are presented as Hamming Corresponding Strategy in Table 7. Results- Comparison. The best feature subset for the Hamming distance measure differed from the best features subset for the Euclidean distance measure. Only three features, mass margin, calcification morphology, and age were present in both feature subsets. The feature set chosen by Euclidean (which in addition included mass density and associated findings) performed slightly better in terms of higher 0.90AUC than the feature set chosen by Hamming (which also included calcification distribution, calcification number, and mass shape). This difference between the 0.90AUC was almost statistically significant ($p=0.06$), while the overall AUCs were comparable (0.79 and 0.80). CBR system with the Euclidean distance measure did perform better at the 95% sensitivity level, but the performance for the two distance measures was similar at the 98% sensitivity level.

As can be seen in Figure 9, the main difference in performance occurs below the 98% sensitivity level, with Euclidean outperforming Hamming. We are unable to say that either distance measure is better for all performance criteria; overall, they seem comparable in performance. It is interesting to examine the profile of malignant cases misclassified by the CBR at high sensitivities. At 98% sensitivity, we are allowing 11 (2%) of the malignancies to be misclassified. Examination of the malignant cases misclassified by Hamming Highest Consistent CBR at 98% sensitivity revealed that the majority of these misclassified cases were mass cases (10 out of 11). One mass case also had calcifications, and another mass case also had a post surgical scar. The eleventh case had calcification findings only. Similar analysis of the malignant cases misclassified by Highest Consistent Euclidean CBR at 98%

36

sensitivity revealed that 7 out of the 11 cases had masses only, 2 had calcifications only, and 2 had focal asymmetric densities only. Therefore, malignant mass cases were misclassified more often than the malignant calcifications. This result is somewhat surprising, since the number of mass and calcification cases in both databases is roughly the same, as well as the number of malignancies in each subset. This could suggest that it might be very advantageous to split the database based on lesion type, and thus improve the performance of the CBR system.

It is important to note that there were only 4 malignant cases that were misclassified by both CBR with Hamming and by CBR with Euclidean at 98% sensitivity. All of the four cases were mass cases with no other findings. All four had a well-circumscribed, isodense mass with either round or oval shape. Further examination of the information that was not given to the CBR about the cases revealed that two of the cases exhibited change from a previous examination. No additional information was available about the third case. The small number of malignant cases misclassified by both distance measures suggests that combining the two classifiers could improve performance.

We have also examined the profile of benign cases potentially spared from biopsy by the CBR. Examination of the benign cases potentially spared from biopsy by Hamming at 98% sensitivity revealed that the vast majority of them (99%) were mass cases. A very small number of cases also had calcifications (0.04%), and other findings (0.03%). Among the mass cases, the predominantly occurring mass margin finding was "well circumscribed," the most often occurring density was "isodense", and the most often occurring shape was "oval." A similar examination of the cases spared by Euclidean at 98% sensitivity showed that 86% were mass cases, and had a similar case profile to cases spared by Hamming. These results indicate further the need to split the database based on lesion type, since most of the calcifications were classified as malignant, while the cases classified as benign

consisted mostly of mass cases. About 72% of the cases spared by Euclidean were also spared by Hamming. The cases spared from biopsy by the CBR matched rather well to the "likely benign" assessment assigned by mammographers. This "gut assessment" ranged from 1 ("benign") to 5 ("malignant"). Please note that this "gut assessment" does not correspond exactly to the BI-RADS ™ assessment. About 18.5% of the cases spared from biopsy by both distance measures had a "benign" mammographers' assessment (category 1), 60% had a "likely benign" assessment (category 2), 18% had a "indeterminate" assessment (category 3), and only 1.5% had a "likely malignant" assessment (category 4). There were no category 5 lesions ("malignant") in the cases spared by both distance measures, and 2% of the cases had no assessment. Our results indicate that the CBR had chosen cases that would fit well with the approach of short-term follow-up, since 96% of the cases had a mammographers' "gut assessment" ranging from 1(benign) to 3 (indeterminate).

Therefore, the CBR presents a potentially useful tool for the classification of mammographic lesions, by recommending short-term follow up for probably benign lesions that is in agreement with the final biopsy result as well as the mammographers' intuition. Although, to our knowledge, the comparison of CBR for breast biopsy decisions to other classifiers has not been published in the literature, the performance of two other classifiers (neural networks and support vector machines, SVMs) has been studied on similar data. In a paper by Markey et al,[41] feed-forward back propagation neural networks were trained and tested on 1453 biopsy cases, and the data set used in that study was almost identical to ours. They report an AUC=0.82±0.01, and $_{0.90}$AUC=0.30±0.03. They also performed linear discriminant analysis, resulting in AUC=0.80±0.01, and $_{0.90}$AUC=0.28±0.03. Compared to the CBR (Euclidean AUC=0.80±0.01, $_{0.90}$AUC=0.37±0.03), the CBR seems to perform similarly in terms of AUC, and better in terms of $_{0.90}$AUC. In a recent paper by Land et al,[80] SVMs were used on 500 breast biopsy cases, which had also been analyzed using neural networks.[28] The SVMs performed similarly to

38

neural networks in terms of AUC (0.85±0.047 for SVMs and 0.86±0.02 for neural networks). While no

0.90AUC was reported, the SVMs seemed to perform better in terms of specificity at high sensitivities. However, the two classifiers used different validation techniques and evaluated specificities at disparate sensitivities, therefore no direct performance comparison at high sensitivities can be made. The main advantage of CBR compared to the aforementioned classifiers is the natural ability of the CBR to explain the reasoning behind the final decision in medical context, while the other classifiers often behave as a "black box." Furthermore, the CBR can supply this additional medical information without sacrifices in performance.

Reportable Outcomes

All years:

Peer-reviewed manuscripts

1. Floyd C.E., Jr.,Lo J.Y., Tourassi G.D., Breast Biopsy: Case-Based Reasoning Computer-Aid Using Mammography Findings for the Decision to Biopsy, American Journal of Roentgenology (AJR) 175:1-6, 2000.

2. A.O. Bilska-Wolak, C.E. Floyd Jr., "Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with BI-RADS ™ lexicon." Med. Phys (in press).

3. A.O. Bilska-Wolak, C.E. Floyd Jr., Joseph Y. Lo, "Application of likelihood ratio to classification of mammographic masses; performance comparison to case •based reasoning." Med. Phys. (submitted).

Conference Proceedings

1. Floyd CE, Jr, Lo JY, Tourassi, GD, "Case-Based Reasoning as a Computer Aid to Diagnosis," Medical Imaging 1999: Image Processing, Hanson KM, Ed., *Proc. SPIE*, 3661:486-489, 1999.

2. A.O. Bilska, C.E. Floyd, Jr, "Investigating different similarity measures for a case-based reasoning classifier to predict breast cancer," SPIE Vol. 4322, p. 1862-1866, 2001.

3. A.O. Bilska-Wolak, C.E. Floyd, Jr, "Breast biopsy prediction using a case-based reasoning classifier for masses versus calcifications." SPIE Vol. 4684, p. 661-665, 2002.

4. A.O. Bilska-Wolak, C.E. Floyd, Jr., Joseph Y. Lo, " Prediction of breast biopsy outcome using a likelihood ratio classifier and biopsy cases from two medical centers." SPIE 2003, (accepted).

Presentations and abstracts

1. Floyd CE Jr., Lo JY, Baker JA, Kornguth PJ Multi-Institution Evaluation of Case-Based Reasoning for Breast Cancer Prediction. *Radiolog* 213(P), 334 1999

2. A.O. Bilska-Wolak, C.E. Floyd, Jr., Joseph Y. Lo, "Computer-assisted prediction of breast biopsy results using radiological data and a likelihood ratio detector." RSNA 2002 (accepted).

Database developed for BIRADS findings of cases referred to biopsy.

Funding applied for from NIH through the R01 mechanism June 2001.

Total:
Two peer-reviewed manuscripts accepted
One peer-reviewed manuscript submitted
Four conference presentations and proceedings
Two conference presentations and abstracts

For a total of 9 publications

Conclusion

In conclusion, the database was analyzed and the distribution of several features was reported, non-parametric evaluation techniques were explored and found to be more appropriate than parametric techniques, the performance of the CBR classifier was examined under variations of several of the key components of the system. The performance was evaluated for different sets of test data and database data from different institutions. Differences in performance were observed. Performance was evaluated under different sets of input findings and an optimal set was selected. Performance was evaluated under different implementations of the Hamming distance criteria under different cutoff distances. Differences were observed and an optimal cutoff was discovered. These interim results suggest that the current study plan is appropriate and that the CBR approach can be developed into a clinically usable decision tool.

REFERENCES

1   L. Garfinkel, C. C. Boring, and C. W. Heath, "Changing trends: an overview of breast cancer incidence and mortality," Cancer **74**, 222-227 (1994).

2   S. Shapiro, "Screening: assessment of current studies," Cancer **74**, 231-238 (1994).

3   A. L. M. Verbeek, J. H. C. L. Hendriks, R. Holland, M. Mravunac, F. Sturmans, and N. E. Day, "Reduction of breast cancer mortality through mass screening with modern mammography," Lancet **1**, 1222-1224 (1984).

4   D. D. Adler, and M. A. Helvie, "Mammographic biopsy recommendations," Current Opinion in Radiology **4**, 123-129 (1992).

5   D. B. Kopans, "The positive predictive value of mammography," AJR Am J Roentgenol **158**, 521-526 (1992).

6   S. Ciatto, L. Cataliotti, and V. Distante, "Nonpalpable lesions detected with mammography: review of 512 consecutive cases," Radiology **165**, 99-102 (1987).

7   F. M. Hall, "Screening mammography - potential problems on the horizon," NEJM **314**, 53-55 (1986).

8   F. M. Hall, J. M. Storella, D. Z. Silverstone, and G. Wyshak, "Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography," Radiology **167**, 353-358 (1988).

9   G. F. Schwartz, D. L. Carter, E. F. Conant, F. H. Gannon, G. C. Finkel, and S. A. Feig, "Mammographically detected breast cancer: nonpalpable is not a synonym for inconsequential," Cancer **73**, 1660-1665 (1994).

10   M. A. Helvie, D. M. Ikeda, and D. D. Adler, "Localization and needle aspiration of breast lesions: complications in 370 cases," AJR Am J Roentgenol **157**, 711-714 (1991).

11   J. M. Dixon, and T. G. John, "Morbidity after breast biopsy for benign disease in a screened population," Lancet **1**, 128 (1992).